Processo seletivo | Engenharia de Dados

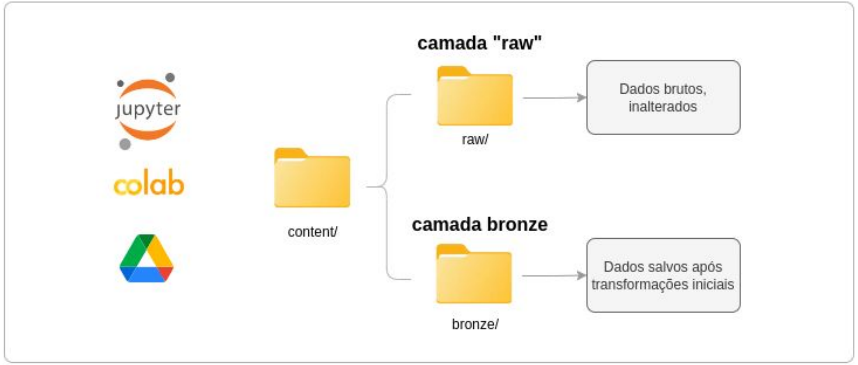
Gabriela Malaspina

Case “Human activity recognition”

21 de março de 2024

Arquitetura da solução

Arquitetura



Etapas

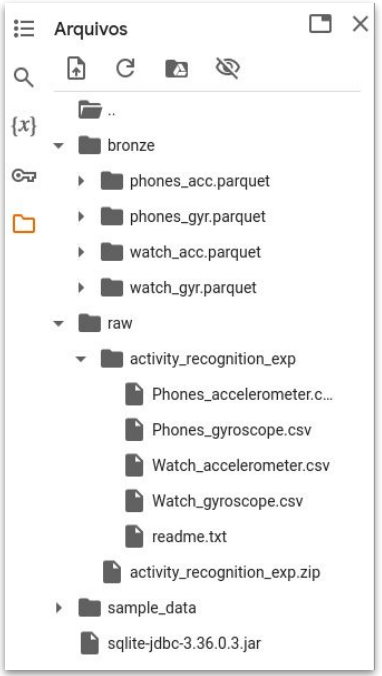
Download (.zip) e extração dos arquivos (.csv) na camada “raw”

1. Leitura (.csv), criação dos dataframes do PySpark e definição manual dos tipos de dados
2. Escrita dos dados no formato .parquet, para otimização de desempenho

2. Análises iniciais e limpeza dos dados; escrita dos dados limpos

3. Criação de novas tabelas com informações dos usuários

Estrutura



Etapas e resultados

Etapa 1

Dataframes PySpark com schemas estruturados manualmente, com base na documentação.

```
# Verificação do Schema estruturado
phones_acc.printSchema()
```

```
root
 |-- Index: integer (nullable = true)
 |-- Arrival_Time: timestamp (nullable = true)
 |-- Creation_Time: timestamp (nullable = true)
 |-- x: float (nullable = true)
 |-- y: float (nullable = true)
 |-- z: float (nullable = true)
 |-- User: string (nullable = true)
 |-- Model: string (nullable = true)
 |-- gt: string (nullable = true)
```

```
# Acelerômetro de celulares
phones_acc.show(5)
# Giroscópio de celulares
phones_gyr.show(5)
# Acelerômetro de relógios
watch_acc.show(5)
# Giroscópio de relógios
watch_gyr.show(5)
```

```
+-----+-------------------+-------------------+----------+----------+--------+----+------+--------+-----+
|Index|       Arrival_Time|      Creation_Time|         x|         y|       z|User| Model|  Device|   gt|
+-----+-------------------+-------------------+----------+----------+--------+----+------+--------+-----+
|    0|2015-02-23 13:03:...|2015-02-23 13:03:...| -5.958191| 0.6880646|8.135345|   a|nexus4|nexus4_1|stand|
|    1|2015-02-23 13:03:...|2015-02-23 13:03:...|  -5.95224| 0.6702118|8.136536|   a|nexus4|nexus4_1|stand|
|    2|2015-02-23 13:03:...|2015-02-23 13:03:...|-5.9950867| 0.6535492|8.204376|   a|nexus4|nexus4_1|stand|
|    3|2015-02-23 13:03:...|2015-02-23 13:03:...|-5.9427185| 0.6761627|8.128204|   a|nexus4|nexus4_1|stand|
|    4|2015-02-23 13:03:...|2015-02-23 13:03:...| -5.991516|0.64164734|8.135345|   a|nexus4|nexus4_1|stand|
+-----+-------------------+-------------------+----------+----------+--------+----+------+--------+-----+
only showing top 5 rows

+-----+-------------------+-------------------+-----------+-------------+------------+----+------+--------+-----+
|Index|       Arrival_Time|      Creation_Time|          x|            y|           z|User| Model|  Device|   gt|
+-----+-------------------+-------------------+-----------+-------------+------------+----+------+--------+-----+
|    0|2015-02-23 13:03:...|2015-02-23 13:03:...|0.013748169|-6.2561035E-4|-0.023376465|   a|nexus4|nexus4_1|stand|
|    1|2015-02-23 13:03:...|2015-02-23 13:03:...|0.014816284|-0.0016937256| -0.02230835|   a|nexus4|nexus4_1|stand|
|    2|2015-02-23 13:03:...|2015-02-23 13:03:...|  0.0158844|-0.0016937256|-0.021240234|   a|nexus4|nexus4_1|stand|
|    3|2015-02-23 13:03:...|2015-02-23 13:03:...|0.016952515| -0.003829956| -0.02017212|   a|nexus4|nexus4_1|stand|
|    4|2015-02-23 13:03:...|2015-02-23 13:03:...|  0.0158844|-0.0070343018| -0.02017212|   a|nexus4|nexus4_1|stand|
+-----+-------------------+-------------------+-----------+-------------+------------+----+------+--------+-----+
only showing top 5 rows

+-----+-------------------+-------------------+----------+----------+----------+----+-----+------+-----+
|Index|       Arrival_Time|      Creation_Time|         x|         y|         z|User|Model|Device|   gt|
+-----+-------------------+-------------------+----------+----------+----------+----+-----+------+-----+
|    0|2015-02-23 13:03:...|1970-01-01 07:45:...| -0.5650316| -9.572019|-0.61411273|   a| gear|gear_1|stand|
|    1|2015-02-23 13:03:...|1970-01-01 07:45:...|-0.83258367| -9.713276|-0.60693014|   a| gear|gear_1|stand|
|    2|2015-02-23 13:03:...|1970-01-01 07:45:...| -1.0181342| -9.935339|-0.54408234|   a| gear|gear_1|stand|
|    3|2015-02-23 13:03:...|1970-01-01 07:45:...| -1.2228385|-10.142437| -0.5662287|   a| gear|gear_1|stand|
|    4|2015-02-23 13:03:...|1970-01-01 07:45:...| -1.5771804|-10.480618|-0.40282443|   a| gear|gear_1|stand|
+-----+-------------------+-------------------+----------+----------+----------+----+-----+------+-----+
only showing top 5 rows

+-----+-------------------+-------------------+-----------+-------------+------------+----+------+--------+-----+
|Index|       Arrival_Time|      Creation_Time|          x|            y|           z|User| Model|  Device|   gt|
+-----+-------------------+-------------------+-----------+-------------+------------+----+------+--------+-----+
|    0|2015-02-23 13:03:...|2015-02-23 13:03:...|0.013748169|-6.2561035E-4|-0.023376465|   a|nexus4|nexus4_1|stand|
|    1|2015-02-23 13:03:...|2015-02-23 13:03:...|0.014816284|-0.0016937256| -0.02230835|   a|nexus4|nexus4_1|stand|
|    2|2015-02-23 13:03:...|2015-02-23 13:03:...|  0.0158844|-0.0016937256|-0.021240234|   a|nexus4|nexus4_1|stand|
|    3|2015-02-23 13:03:...|2015-02-23 13:03:...|0.016952515| -0.003829956| -0.02017212|   a|nexus4|nexus4_1|stand|
|    4|2015-02-23 13:03:...|2015-02-23 13:03:...|  0.0158844|-0.0070343018| -0.02017212|   a|nexus4|nexus4_1|stand|
+-----+-------------------+-------------------+-----------+-------------+------------+----+------+--------+-----+
only showing top 5 rows
```

Etapas e resultados

Etapa 2

Análises iniciais, problemas encontrados e possíveis tratativas.

Análises iniciais:

- Verificação da volumetria;
- Contagem de dados nulos por coluna;

```
Quantidade de linhas "Acelerômetro de celulares": 13062475
Quantidade de linhas "Giroscópio de celulares": 13932632
Quantidade de linhas "Acelerômetro de relógios": 3540962
Quantidade de linhas "Giroscópio de relógios": 13932632
```

```
+-----+------------+-------------+---+---+---+----+-----+------+---+
|Index|Arrival_Time|Creation_Time|  x|  y|  z|User|Model|Device| gt|
+-----+------------+-------------+---+---+---+----+-----+------+---+
|    0|           0|            0|  0|  0|  0|   0|    0|     0|  0|
+-----+------------+-------------+---+---+---+----+-----+------+---+
```

Desafios encontrados:

- Variação da volumetria:
  - Dados ausentes ou nulos;
  - Processamento anterior;
  - Erros nos dados;
  - Heterogeneidade das amostras.
- Volumetria elevada;
- Contexto sobre o assunto.
- Datas incorretas

Proposta:

Investigação dos dados e revisão das etapas anteriores;

Alinhamento e validação com time de negócio;

Utilização de formato .parquet;

Documentação, pesquisa e alinhamento com negócio.

Alinhamento e validação com time de negócio ou cálculo da correlação entre datas para substituição dos valores incorretos

Etapas e resultados

Etapa 3

Informações sumarizadas por usuário.

Database unificada:

```
user_database.show(5)
```

```
+------+----+----------+------------+----+----------+
| Index|User|     Model|      Device|  gt|    Source|
+------+----+----------+------------+----+----------+
|107066|   a|samsungold|samsungold_2|bike|phones_acc|
|107067|   a|samsungold|samsungold_2|bike|phones_acc|
|107068|   a|samsungold|samsungold_2|bike|phones_acc|
|107069|   a|samsungold|samsungold_2|bike|phones_acc|
|107070|   a|samsungold|samsungold_2|bike|phones_acc|
+------+----+----------+------------+----+----------+
only showing top 5 rows
```

Resumo dos dados:

```
Base unificada:
 Usuários distintos: 9
 Total de registros: 52249900
```

Registros distintos por usuário:

```
+----+---------------+
|User|Total_Registros|
+----+---------------+
|   a|        5450080|
|   b|        6195072|
|   c|        5307192|
|   d|        5348496|
|   e|        6459996|
|   f|        5538124|
|   g|        6350788|
|   h|        5369204|
|   i|        6230948|
+----+---------------+
```

Dispositivos distintos por usuário:

```
+----+--------------------+-------------+
|User|       Device_Models|Total_Modelos|
+----+--------------------+-------------+
|   a|[nexus4, s3mini, ...|            4|
|   b|[nexus4, s3mini, ...|            4|
|   c|[nexus4, s3mini, ...|            4|
|   d|[nexus4, s3mini, ...|            4|
|   e|[nexus4, s3mini, ...|            4|
|   f|[nexus4, s3mini, ...|            4|
|   g|[nexus4, s3mini, ...|            4|
|   h|[nexus4, s3mini, ...|            4|
|   i|[nexus4, s3mini, ...|            4|
+----+--------------------+-------------+
```

Processo seletivo | Engenharia de Dados

Gabriela Malaspina

Obrigada!

21 de março de 2024